Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models

Robert Pincus[1], Crispian P. Batstone[1], Robert J. Patrick Hofmann[1],
Karl E. Taylor[2] and Peter J. Glecker[2]

[1]*University of Colorado and NOAA Earth System Research Laboratory, Boulder, Colorado*
[2]*Atmospheric Sciences Division, Lawerence Livermore National Laboratory, Livermore, California*

January, 2008

Revised for *J. Geophys. Res.*

Corresponding Author Address:
Robert Pincus
Cooperative Institute for Research in Environmental Sciences,
University of Colorado/NOAA Earth Systems Research Laboratory
Physical Sciences Division
325 Broadway, R/PSD1
Boulder CO 80305-3337
Robert.Pincus@colorado.edu

ABSTRACT

This paper describes a set of metrics for evaluating the simulation of clouds, radiation, and precipitation in the present-day climate. As with the skill scores used to measure the accuracy of short-term weather forecasts, these metrics are low-order statistical measures of agreement with relevant, well-observed physical quantities. The metrics encompass five statistical summaries computed for five physical quantities (longwave, shortwave, and net cloud radiative effect, projected cloud fraction, and surface precipitation rate) over the global climatological annual cycle. Agreement is measured against two independent observational datasets.

The metrics are computed for the models that participated in the Coupled Model Intercomparison Project phase 3, which formed the basis for the Fourth Assessment of the IPCC. Model skill does not depend strongly on the dataset used for verification, indicating that observational uncertainty does not limit the ability to assess model simulations of these fields. No individual model excels in all scores though the "IPCC mean model," constructed by averaging the fields produced by all the CMIP models, performs particularly well across the board. This skill is due primarily to the individual model errors being distributed on both sides of the observations, and to a lesser degree to the models having greater skill at simulating large-scale features than those near the grid scale. No measure of model skill considered here is a good predictor of the strength of cloud feedbacks under climate change.

The model climatologies, observational data sets, and metric scores are available on-line.

1 Assessing the skill of weather forecasts and climate projections

Global numerical weather forecasts have been made operationally since 1979 and have improved steadily since their advent [*Kalnay et al.*, 1998; *Simmons and Hollingsworth*, 2002]. It is possible to trace the quantitative increase in skill over time because forecasts have been evaluated against observations in a consistent manner for decades. The World Meteorological Organization, for example, defines a "standard verification system" (SVS) which is a set of low-order statistics measuring forecast skill that operational forecasting centers compute monthly and share with one another [*World Meteorological Association*, 1992]. Skill is measured against both raw observations (typically radiosondes) and against the analysis produced after the fact for the forecast time.

Weather forecast assessment is primarily focused on the large-scale flow: only temperatures, winds, sea level pressure, and geopotential height are part of the standard verification system. In particular, no information is formally exchanged between centers about forecasts of clouds, radiation, or precipitation (though centers may verify precipitation forecasts internally). Clouds and radiation are neglected in part because verification is difficult: these fields are much more spatially and temporally variable than winds, temperature, and pressure, so measurements made at individual points are less representative of the grid-column mean produced by the forecast. Analyses are not useful for verification, either, because observations of clouds and broadband radiation are not used by operational assimilation systems, so that cloud- and radiation-related quantities are only loosely constrained in the resulting analyses. In addition, though the distribution of clouds, radiation, and precipitation may be of importance for specific applications, it has a minor impact on the short- to medium-term evolution of the flow on which forecast skill scores are based.

Unlike weather forecasts, the climate models used to make long-term projections have not been subject to uniform assessment over time. Climate model evaluation is, in some ways, more difficult than assessing the skill of short-term weather forecasts because climate models solve a boundary value problem as opposed to the initial value problem posed in weather forecasting. This blurs the association between time in the model and time in nature, so model forecasts can't be compared with observations on a day-to-day or month-to-month basis. Furthermore, climate models are primarily used to make projections over very long time scales (decades to centuries), and these projections can not be directly assessed until that time has passed.

In lieu of direct assessment of long-term trends, climate models are evaluated according to their ability to simulate present-day conditions and the historical record. Evaluation typically includes comparisons against observations of the mean climate and of variability at various forced (annual, diurnal) and unforced (El Niño-Southern Oscillation, Madden-Julian Oscillation) time scales, and may include the response to forcings such as volcanic eruptions and changes in atmospheric composition. (See, as one example, the papers describing the current version of the coupled climate model CM 2.1 produced by the Geophysical Fluid Dynamics Laboratory, including *GFDL GAMDT* [2004] and *Wittenberg et al.* [2006].) Attempts to identify which aspects (if any) of the current climate are the best predictors of climate sensitivity or related quantities have so far been unsuccessful, so a model's skill in simulating the present day climate may not reflect accuracy in long-term projections. Nonetheless, it seems unlikely that a model that does a poor job simulating the current climate will somehow produce credible long-term projections. This is one motivation for developing a suite of metrics for evaluating a model's climatology (e.g. *Gleckler et al.*, 2007, hereafter GTD2007).

Clouds strongly modulate the long-term evolution of the atmosphere, and cloud feedbacks on the climate system have remained the single largest source of uncertainty in climate projections since the Intergovernmental Panel on Climate Change (IPCC; see http://www.ipcc.ch) began issuing Assessment Reports in 1990 (compare the discussions of uncertainty in the First Assessment report [*Houghton et al.*, 1990] and the Fourth [available from http://ipcc-wg1.ucar.edu/wg1/wg1-report.html]). For this reason there is far greater motivation for including the clouds in the evaluation of climate models than in short-term forecast models. There are, however, no standard metrics for judging model skill in simulating present-day cloudiness or related quantities such as rainfall and radiation.

This paper proposes metrics for evaluating global simulations of radiation, clouds, and precipitation, focusing on measures most relevant for evaluating multi-year simulations. The next section describes the metrics in detail, including choices regarding the quantities, domain, and summary measures, while Section 3 reports the scores for the current generation of climate models and several other models that might be expected to perform substantially better. We will demonstrate that the best agreement with the present-day distribution of clouds and related quantities comes from averaging over all the available models, and in Section 4 we explore the mechanisms that lead to this result. As we describe in Section 5, we find no relationship between skill in simu-

lating the present-day climate and the cloud feedback parameter that plays an important role in determining climate sensitivity.

2 Metrics for assessing global simulations of clouds, radiation, and precipitation

In developing metrics for evaluating clouds in climate models we have been guided by the desire to stay as close as possible to the WMO standard verification system. Metrics are defined by four choices: the physical parameter being evaluated; the set of observations against which the models are evaluated; the space and time domain over which statistics are computed; and the statistical measure used. We seek to evaluate the simulation of quantities that are both well-observed and relevant to climate change projections.

a. Quantities and observational data sets

We evaluate five quantities: total cloud fraction, surface precipitation rate, and three measures of the cloud radiative effect at the top of the atmosphere (clear-sky flux minus total flux for longwave [LW], shortwave [SW], and the net radiation at TOA). Global observational estimates of each of these quantities are available from two or more independent programs. Cloud radiative effect is included to provide a coarse but integrated measure of the distribution of cloud properties in space and time. For comparison we also compute scores for the TOA all-sky LW, SW and net fluxes, but metrics based on these fluxes are less indicative of model skill for several reasons. Most importantly, the spatial and temporal patterns of TOA flux are strongly constrained by the pattern of insolation, including the strong equator-to-pole gradient and the seasonal cycle, and all models compute the insolation substantially correctly. (Insolation affects the outgoing longwave flux through the temperature response.) Secondarily, all climate models are tuned until energy balance is achieved at the TOA in the global, annual mean, and many [see, e.g. *GFDL GAMDT* 2004] tune so that the mean SW and LW fluxes at the TOA match the ERBE observations, which makes comparison with those fluxes less independent

Table 1 lists the primary and secondary datasets used for each quantity, along with the epoch for which observations are available. We assume that each set of observations is sufficient to describe a time-stationary climatology, though the length of the observational records varies by

parameter and data source. As we will show in the next section, the primary and secondary observational estimates of each field are, in every case, in better agreement with each other than the models are with either data source, so our results are not strongly influenced by this assumption.

b. Temporal and spatial domain

Following the standard verification system, we interpolate all quantities onto a uniform 2.5 x 2.5 degree grid. We compute the mean annual cycle by separately averaging each month over the observational record to obtain 12 maps, one for each month, of each quantity. We have not decomposed the globe into subdomains as the WMO Standard Verification System does, but this might be easily done.

c. Statistical summaries

The standard verification system used by weather forecasting centers comprises root-mean-square (RMS) error, mean error (bias), and anomaly correlation for each variable, computed separately for each forecast and averaged over each month. By analogy to these scores we compute five statistical summaries: RMS error e, mean bias $\bar{e}$, centered RMS error e', the ratio of the standard deviations s, and the correlation r, defined so that bias is positive when model values exceed those observed and $s > 1$ when model values are more variable than observed. (We replace anomaly correlation with correlation because the former requires the climatological value for the space/time domain to be removed but, in our case, climatology is exactly the signal we seek to evaluate.) The five quantities are related by

$$\begin{aligned} e^2 &= \bar{e}^2 + e'^2 \\ &= \bar{e}^2 + \sigma_o^2(1 + s^2 - 2sr) \end{aligned} \tag{1}$$

where σ_o refers to the standard deviation of the observed field in the time/space domain. The geometric relationships in (1) form the basis of the "Taylor diagrams" [*Taylor*, 2001] used to visualize results in the next section. We compute statistics over the twelve monthly mean maps, weighting each column by its surface area, to assess the simulation of both the annual cycle and spatial variability (c.f. *GDT2007*).

3 How well does the current generation of climate models simulate clouds, radiation, and precipitation?

a. Models assessed in this study

We compute statistics for the set of climate models that participated in the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset used to prepare the IPCC Fourth Assessment (IPCC AR4; see http://ipcc-wg1.ucar.edu/wg1/wg1-report.html). Monthly mean values were obtained from the archives maintained by the Project for Climate Model Diagnosis and Intercomparison (PCMDI) and described by *Meehl et al.* [2007]. We evaluate the models' performance from 1979-1999 in two sets of runs: one in which the atmospheric models are coupled to dynamical ocean models that compute the sea surface temperature (20th century runs, for which 22 models provided output) and one in which the sea surface temperature is specified (AMIP runs, 12 of the 22 models). The models are listed in Table 2 and described in more detail on the PCMDI web site (http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php). Many models submitted ensembles of simulations; for these we use the ensemble mean averaged over each month. (Results in Section 4 indicate that model performance does not depend strongly on ensemble size.)We have excluded the BCC CM1 because, as of this writing, the fields in the archives were so different from the observations as to dominate our results. Not all models provided data for every field.

In addition to the models participating in CMIP3/AR4, we compute statistics for four models that might be expected to behave substantially differently. One is the "super-parameterized" version of the NCAR Community Atmosphere Model [*Khairoutdinov et al.*, 2005] in which the physical parameterizations are replaced by a copy of a two-dimensional fine-scale cloud resolving model in each grid cell. The Super-CAM has been run for 13 years (1986-1999) using AMIP-specified sea surface temperatures [*Khairoutdinov et al.*, 2007]. The second is the "IPCC mean model" computed by averaging the monthly mean fields provided by each model (using the single ensemble mean for models that provided ensembles), computed separately for the AMIP and 20th century ensembles. We also include 12-hour forecasts created as part of the 40-year reanalysis product by the European Centre for Medium-Range Weather Forecasts (ERA-40, see *Uppala et al.*, 2005). Over much of the globe wind, temperature, and humidity fields in these forecasts are

tightly constrained by observations during the analysis cycle. These fields are therefore expected to be much closer to those observed than are the free-running climate models, so errors in the cloud, radiation, and precipitation fields can be attributed to errors in the underlying cloud or boundary layer parameterizations to a much greater extent than in any of the other models. Finally, we include a run of a current version (Cycle 32R1) of the ECMWF forecast model run from 1 Dec 1991 - 1 Dec 2001 using specified sea surface temperatures (i.e. as an AMIP simulation). This model has substantially higher resolution than most climate models (T159 with 91 levels) but has been developed for weather forecasting applications rather than climate projections.

As with the observations, we compute mean monthly climatologies of radiative fluxes, cloud radiative effect, cloud fraction, and surface precipitation rate by averaging each month over the twenty-year period (or whatever period is available).

b. Skill measures

Figure 1 provides a qualitative overview of the relative skill of each model in simulating the present-day annual cycle and spatial distribution of clouds, radiation, and surface precipitation. Each row of this "portrait diagram" (see *GTD2007*) corresponds to a skill measure (e.g. RMS errors in top-of-atmosphere cloud radiative effect) and each column to a model run; AMIP and 20th century runs of the same model have been grouped together. The upper left and lower right triangles in each square show performance as measured against the primary and secondary observations respectively. We convert the five statistics we have calculated to quantities that increase monotonically with model error by showing e, e', $|\bar{e}|$, $1 - r$, $|1 - s|$. These error measures are then expressed as a fractional deviation from the mean value of that error measure, where the mean is computed across all the CMIP3 models (including the mean model). That is, for a given error measure E computed by evaluating field f against model m using observational data set r the fractional error is defined by

$$E'_{mfr} = \frac{E_{mfr} - \overline{E_{fr}}}{\overline{E_{fr}}} \tag{2}$$

Normalization by $\overline{E_{fr}}$ is possible because mean errors are significantly larger than 0 in the set of metrics considered here. In Figure 1, values of $E'_{mfr} < 0$ (i.e. models with smaller errors than the mean error) are shaded in blue and values of $E'_{mfr} > 0$ in red. The black horizontal line separates

the metrics computed for cloud fraction, precipitation rate, and cloud radiative effect from those computed for top-of-atmosphere fluxes.

Several conclusions regarding the CMIP3 models can be drawn from Figure 1. First, although some models agree well with observations in many aspects of the simulation, all models have their weak areas, as indicated by areas of reddish color in every column. Secondly, in almost all cases, the IPCC mean model is closer to the observations than any of the individual models. GTD07 note both of these behaviors across a much wider range of metrics in the 20th century simulations. In most cases the relative performance of individual models across this set of metrics does not depend strongly on the verification data set (i.e. the two triangles in each box are typically close in shade), implying that all simulations of clouds, radiation, and precipitation differ markedly from the observations regardless of the data set used to define climatology. Finally, we note that a model's relative performance with respect to any given metric is not especially sensitive to whether the model is run with specified sea surface temperatures or using a dynamic ocean and 20th century forcings.

The performance of the alternative models provide a useful point of comparison to the results from the CMIP3 climate models. The results from ECMWF reflect, in part, the effects of model development over time. ERA-40 was produced with a model introduced in Oct, 1999, using short forecasts from analyses; C32R1 was introduced in September 2006 and, for the data provided here, was run in a mode similar to climate models, so that day-to-day dynamics and thermodynamics are not constrained by observations. Nonetheless, errors in most cloud, radiation, and precipitation quantities are substantially smaller in C32R1 than in ERA-40. (Errors in cloud fraction are comparable.) It is also possible that some errors in ERA-40 arise from imbalances in the analyses that the model removes in the first few hours of the forecast. The super-parameterized CAM, in contrast, does not perform particularly well on these metrics for clouds, radiation, and precipitation, despite the fact that many other aspects of simulations with the super-parameterized CAM are improved over the standard version of the model [*Khairoutdinov et al.*, 2007]. This may reflect the relatively short amount of time and limited resources that have been devoted to tuning the model to the present-day climate. We have omitted results for cloud fraction because, unlike every other model in our sample, cloud fraction in the super-parameterized CAM does not affect radiative fluxes, which are calculated using the exact amount of condensate in each column in the cloud-resolving model. Comparisons using the ISCCP simulator [*Klein and Jakob*, 1999] are in

reasonable agreement with observations [*Khairoutdinov et al*., 2007], but these estimates incorporate information about the ISCCP cloud detection method, making them difficult to compare to the rest of the models.

A more quantitative view of the agreement between the observations and model simulations may be obtained from the Taylor diagrams in Figure 2, which show, for each data set, the correlation with the observations and the standard deviation (i.e. the terms r and s that make up the centered RMS error e'^2 as shown in (1)). The centered RMS error is indicated by the distance to the intersection of the dashed line and the x-axis with units and measures identical to the x-axis.We have modified the diagrams to show the bias $\bar{e}$ with symbols whose size can be measured against the radial axis. Metrics for each quantity are computed with respect to the primary data set listed in Table 1, and differences between the primary and alternative observational estimates are shown in red. A data set that agreed perfectly with the observations (i.e. had $s = 1$, $e' = 1$, and $\bar{e} = 0$, and hence $e = 0$) would lie on the horizontal axis where it intersected the dashed line. Models may have the same correlation with observational data set while being poorly correlated with each other, so points which are close to one another on the Taylor diagram should be understood as being equally far from the observational data set rather than necessarily close to one another.

The Taylor diagrams confirm several conclusions drawn from the portrait diagram, namely that the mean model performs better on essentially every count than do any of the individual models that comprise it, and that model simulations of clouds, radiation, and precipitation are not systematically better in either the AMIP or 20th century simulations. The Taylor diagrams also demonstrate that the alternative observational data set is in better agreement with the primary observations than are any of the models. The agreement is notable because, for any given quantity, the two observations may be derived from different instruments during different, even non-overlapping, epochs. This indicates that model simulations disagree with the observational record so much that observational uncertainty does not limit the ability to gauge model improvement.Simulations of top-of-atmosphere radiative fluxes are uniformly in better agreement with observations than simulations of cloud radiative effect for the reasons discussed in Section 2.

Taylor diagrams for the net, longwave, and shortwave fluxes are shown for comparison in Figure 3. As we noted in Section 2, the spatial and temporal structure of these fields is dominated by distribution of insolation, which is well-modeled.

4 Why does the mean model perform so well?

As we noted in the last section, the best agreement in the predictions of clouds, radiation, and precipitation is achieved by the IPCC mean model, which is constructed by averaging each field during each month across all models in the ensemble. GTD07 note the same behavior across a much wider range of physical variables. There are at least two mechanisms that might lead to this result. One possibility is that models do a better job simulating large-scale, slowly varying features than those nearer the grid scale at monthly time resolution. High scores for the mean model fields might then arise because the mean model fields are smoother in space and time than any of the individual models, leaving the large-scale agreement but removing the small-scale errors. The second possibility is that the systematic errors associated with individual models are, to some extent, distributed around no error, so that averaging across a range of models reduces those systematic errors. Here we attempt to distinguish the degree to which each of these mechanisms is responsible for the success of the IPCC mean model by comparing metrics computed for a single model run with those computed a) when the model fields have been smoothed and b) for fields derived from a range of ensembles.

We use a 20th century run from the MRI CGCM ensemble as a baseline since the ensemble is of intermediate size (five members) and, in the mean, performs reasonably well on most metrics. We compare the skill of this single model run to 1) the four other individual runs from the MRI CGCM ensemble; 2) the mean of the MRI CGCM ensemble; 3) the same individual run of the MRI CGCM after spatial smoothing; 4) a set of five member ensembles consisting of our baseline run and four other individual runs chosen at random from all individual runs in the IPCC ensemble; and 5) the IPCC mean model. We compute the spatially-smoothed version of the baseline run (item 3) by replacing the climatological value in each grid cell with the running mean of that value and its four closest neighbors, then iterating the process. We convert metric scores to errors as described in Section 3, then normalize each error by the error in the baseline run (i.e. $\tilde{E}_{mf1} = E_{mf1} / E_{bf1}$, where b indicates the baseline run and 1 denotes the primary observation dataset).

Figure 4 shows the distribution of relative errors in all 25 metrics for each of these four scenarios. Errors are comparable in each of the members of the MRI CGCM ensemble: relative errors across all metrics and all four alternate ensemble members range from 89-119%, with a mean of

99.8%. The ensemble mean, constructed by averaging the monthly map of each quantity across ensemble members before computing the metrics, performs very slightly better (range of 92-107% and mean of 99%). Spatial smoothing improves the metric scores modestly but measurably. The third column of Figure 4 shows the distribution of relative errors when each field has undergone three iterations of spatial smoothing, which results in a mean relative error of 89%. (The mean relative error decreases for each of the first four iterations of smoothing; the value is nearly the same after four or five iterations.) Finally, we compute scores for a set of five multi-model ensembles containing five members each (to match the size of the MRI CGCM ensemble) but containing, in addition to the baseline run, four individual runs chosen randomly from the full set submitted to CMIP3. Excepting the mean bias for shortwave cloud radiative effect and surface precipitation (two scores on which the baseline run does particularly well), the multi-model ensembles are in substantially better agreement with the observations than any version of the single GCM: the mean relative error across all metrics is 67%, nearly as small as the mean relative error of the full IPCC ensemble (63%; see the last column of Figure 4).

We infer that the IPCC mean model's good agreement with observations is due to both spatial smoothing and compensating systematic errors, but that the latter is more important.

5 Relating cloud feedbacks to present-day simulations of clouds, radiation, and precipitation

As we noted in Section 1, the most significant application of climate models is for projections of long-term climate change, for which climate sensitivity is often used as a proxy. Much of the diversity in estimates of climate sensitivity, in turn, arises from cloud-related feedbacks on climate change. In particular, much of the range in climate sensitivity in present-day models can be traced to differences in how shallow clouds in the tropics respond to climate change [*Bony and Dufresne*, 2005]. It would therefore be intriguing if there were a systematic relationship in this set of models between present-day skill in predicting cloud fields and some measure of how clouds respond to a changing climate.

We have searched for an association by linearly regressing each of the metrics shown in Figure 1 with estimates of the cloud feedback parameter made by *Soden and Held* [2006] and find no such relationships. Linear correlation coefficients vary from 0 to a maximum of 0.53, indicat-

ing that none of our metrics for evaluating the fidelity of present-day simulations of clouds and radiation is a good predictor of cloud feedbacks under climate change in this set of models. We have also inspected the resulting scatter plots but find no evidence for any non-linear relationships.

6 Conclusions

Though climate and weather forecasting models are structurally similar (and, in at least one case, even the same code), the culture of assessment in the two communities is quite different. To some degree this reflects the fact that weather forecasts can be verified directly in a way that is impossible for climate projections. Nonetheless, it is striking that weather forecasting centers can point to almost two decades worth of improving forecast skill scores as evidence that investment in forecasting leads to better forecasts, while the climate modeling community has a difficult time making a similar case quantitatively. We suggest that this argues for the routine calculation and dissemination of performance metrics for climate models. Such metrics might make it easier to defuse criticism based, for example, on the fact that the range of climate sensitivity reported in the IPCC assessments has not narrowed over time by demonstrating that the models have gotten better at simulating the present-day climate. As we noted in Section 1, though agreement with present-day observations does not guarantee that projections of climate change will be correct, such agreement seems desirable.

We expect that comprehensive evaluation of the simulation of present clouds, radiation, and precipitation could require more metrics than are presented here. To assist in the development of additional metrics we have made available the composite annual cycles for each model run and for the observational data sets used in this study, along with the metrics themselves. These may be obtained by visiting [web address to be provided]. We expect, however, that some useful metrics can not be computed using only the composite annual cycle. This is perhaps most true for precipitation, since rainfall rate varies so dramatically in space and time and since, in many cases, it is the extreme events which matter most. We expect that there may be some utility in adapting quantitative precipitation forecast skill scores (see, e.g., *Wilks*, 1995), which account for the entire distribution of rainfall rates, to the assessment of climate models; this will necessitate changes in the data archiving strategy at most climate modeling centers.

Figures 1 and 2 show that the two estimates of top-of-atmosphere cloud radiative effect and fluxes we have used (CERES ES-4 "ERBE-like" and ERBE S-4G) are in better agreement with each other than with any of the models (including the mean model), indicating that observational uncertainty does not limit our ability to gauge model performance. This may change, however, with the advent of new CERES data products. These estimates begin with the same measurements as the ES-4 products but use substantially improved algorithms including better scene identification, a wider diversity of angular models, and more accurate time interpolation. At this time it is not possible to use these products to compute cloud radiative effect: only the CERES SRBAVG products are available as yet, and these contain significant amounts of missing values for clear-sky fluxes on a monthly basis because the tests for clear sky are so stringent. Data sets that use additional information to estimate the clear-sky fluxes (i.e. the AVG and SYN data sets) are still in development. When they are released these data are likely to be substantially more accurate than ERBE, so quantities may well differ from the ERBE estimates by amounts as large as the difference between ERBE and the current generation of models. Nonetheless, we anticipate that model performance with respect to radiation will be best judged against these new products when they become available.

Acknowledgements

We thank the modeling groups, PCMDI, and the WCRP Working Group on Coupled Modelling for their roles in making available the WCRP CMIP3 multi-model dataset. Support for this dataset is provided by the Office of Science, U.S. Department of Energy. We appreciate the efforts of Adrian Tompkins and Thomas Jung, who provided data from the ECWMF and of Marat Khairoutdinov, who gave us output from the super-parameterized CAM. Section 4 was motivated by questions from Susan Solomon. We benefitted from the comments of two anonymous referees and from conversations with Christian Jakob, Bill Rossow, Steve Klein, Thomas Reichler, David Richardson, Martin Miller, and Beth Ebert. This research was supported by NASA under award NNG04GK10G, by NSF under ATM-0336702, and by the Office of Science, U.S. Department of Energy under contract DE-AC52-07NA27344.

References

Bony, S., and J.-L. Dufresne (2005), Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, *J. Geophys. Res.*, 32, L20806, doi:10.1029/2005GL023851.

GFDL Global Atmospheric Model Development Team (2004), The new GFDL global atmosphere and land model AM2-LM2: Evaluation with prescribed SST simulations, *J. Climate*, 17, 4641-4673, doi:10.1175/JCLI-3223.1

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2007), Performance metrics for climate models, *J. Geophys. Res.*, 32, to appear.

Houghton, J. T., G. J. Jenkins, and J. J. Ephraums (1990), Scientific Assessment of Climate change – Report of Working Group I, Cambridge University Press, Cambridge, UK.

Kalnay, E., S. J. Lord, and R. D. McPherson (1998), Maturity of operational numerical weather prediction: Medium range, *Bull. Amer. Met. Soc.*, 79, 2753-2769.

Khairoutdinov, M., C. DeMott, and D. Randall (2005), Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes, *J. Atmos. Sci.*, 62, 2136-2154, doi:10.1175/JAS3453.1.

Khairoutdinov, M., C. DeMott, and D. A. Randall (2007), Evaluation of the simulated interannual and sub-seasonal variability in an AMIP-style simulation using the CSU Multiscale Modeling Framework, *J. Climate*, to appear.

Klein, S. A. and C. Jakob, 1999: Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Weather Rev.*, 127**,** 2514-2531.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007): THE WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Amer. Met. Soc.*, 88, 1383-1394, doi:10.1175/BAMS-88-9-1383.

Simmons, A. J., and A. Hollingsworth (2002), Some aspects of the improvement in skill of numerical weather prediction, *Quart. J. Royal Met. Soc.*, 128, 647-677, doi:10.1256/003590002321042135.

Soden, B. J., and I. M. Held (2006), An assessment of climate feedbacks in coupled ocean-atmosphere models, *J. Climate*, 19, 3354-3360, doi:10.1175/JCLI3799.1.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res. - Atmos.*, 106, 7183-7192, 10.1029/2000JD900719.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Quart. J. Royal Met. Soc.*, 612, 2961-3012, 10.1256/qj.04.176.

Wilks, D. S. (1995), Statistical methods in the atmospheric sciences: An introduction. 467 pp. Academic Press.

Wittenberg, A. T., A. Rosati, N. C. Lau, and J. J. Ploshay (2006), GFDL's CM2 global coupled climate models. Part III: Tropical pacific climate and ENSO, *J. Climate*, 19, 698-722, doi:10.1175/JCLI3631.1.

World Meteorological Association (1992), Manual on the Global Data-Processing System, Geneva, Switzerland. Available from http://www.wmo.int/web/www/documents.html.

Table 1: Quantities used to evaluate clouds, radiation, and precipitation in climate model simulations, and the observations against which the models are compared.

Quantity	Observational data set	Epoch	Alternative observations	Epoch
Top-of-atmosphere cloud radiative forcing (LW, SW, net)	CERES ES-4 ("ERBE-like")	Mar 2000 - Dec 2005	ERBE S-4G	Nov 1984 - Feb 1990
Cloud fraction	ISCCP D2	Jul 1983- Dec 2004	MODIS/ Terra Collection 5	Mar 2000 - Dec 2006
Surface precipitation rate	GPCP v2	Jan 1979 - Apr 2005	CMAP	Jan 1979 - June 2002
Top-of-atmosphere radiative flux (LW, SW, net)	CERES ES-4 ("ERBE-like")	Mar 2000 - Dec 2005	ERBE S-4G	Nov 1984 - Feb 1990

Table 2: Models for which metrics have been computed. Where ensembles have been provided the monthly average is computed across all members of the ensemble. The "IPCC mean model" is the average of the monthly climatologies for each model. Where centers have submitted more than one model the institution and country are listed only for the first model.

Institution	Country	Model name	Ensemble size - AMIP	Ensemble size - 20th C
Bjerknes Centre for Climate Research	Norway	BCM 2.0		1
National Center for Atmospheric Research	USA	CCSM 3	1	4
		PCM	1	2
Canadian Centre for Climate Modelling & Analysis	Canada	CGCM 3.1 (T47)		5
Canadian Centre for Climate Modelling & Analysis	Canada	CGCM 3.1(T63)		1
Météo-France / Centre National de Recherches Météorologiques	France	CNRM CM3	1	1
CSIRO Atmospheric Research	Australia	CSIRO Mk3.0		3
Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-OM	3	3
Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA	Germany, Korea	ECHO-G		5
LASG / Institute of Atmospheric Physics	China	FGOALS g1.0	3	3

Table 2: Models for which metrics have been computed. Where ensembles have been provided the monthly average is computed across all members of the ensemble. The "IPCC mean model" is the average of the monthly climatologies for each model. Where centers have submitted more than one model the institution and country are listed only for the first model.

Institution	Country	Model name	Ensemble size - AMIP	Ensemble size - 20th C
NOAA / Geophysical Fluid Dynamics Laboratory	USA	GFDL CM 2.0		3
		GFDL CM 2.1		3
NASA / Goddard Institute for Space Studies	USA	GISS AOM		2
		GISS EH		5
		GISS ER	4	9
Institute for Numerical Mathematics	Russia	INM CM 3.0	1	1
Institut Pierre Simon Laplace	France	IPSL CM 4	6	2
Center for Climate System Research, National Institute for Environmental Studies, and Frontier Research Center for Global Change	Japan	MIROC 3.2 (hires)	1	1
		MIROC 3.2 (medres)	3	3
Meteorological Research Institute	Japan	MRI CGCM 2.3.2	1	5
Hadley Centre for Climate Prediction and Research / Met Office	UK	UKMO HadCM3		2

Table 2: Models for which metrics have been computed. Where ensembles have been provided the monthly average is computed across all members of the ensemble. The "IPCC mean model" is the average of the monthly climatologies for each model. Where centers have submitted more than one model the institution and country are listed only for the first model.

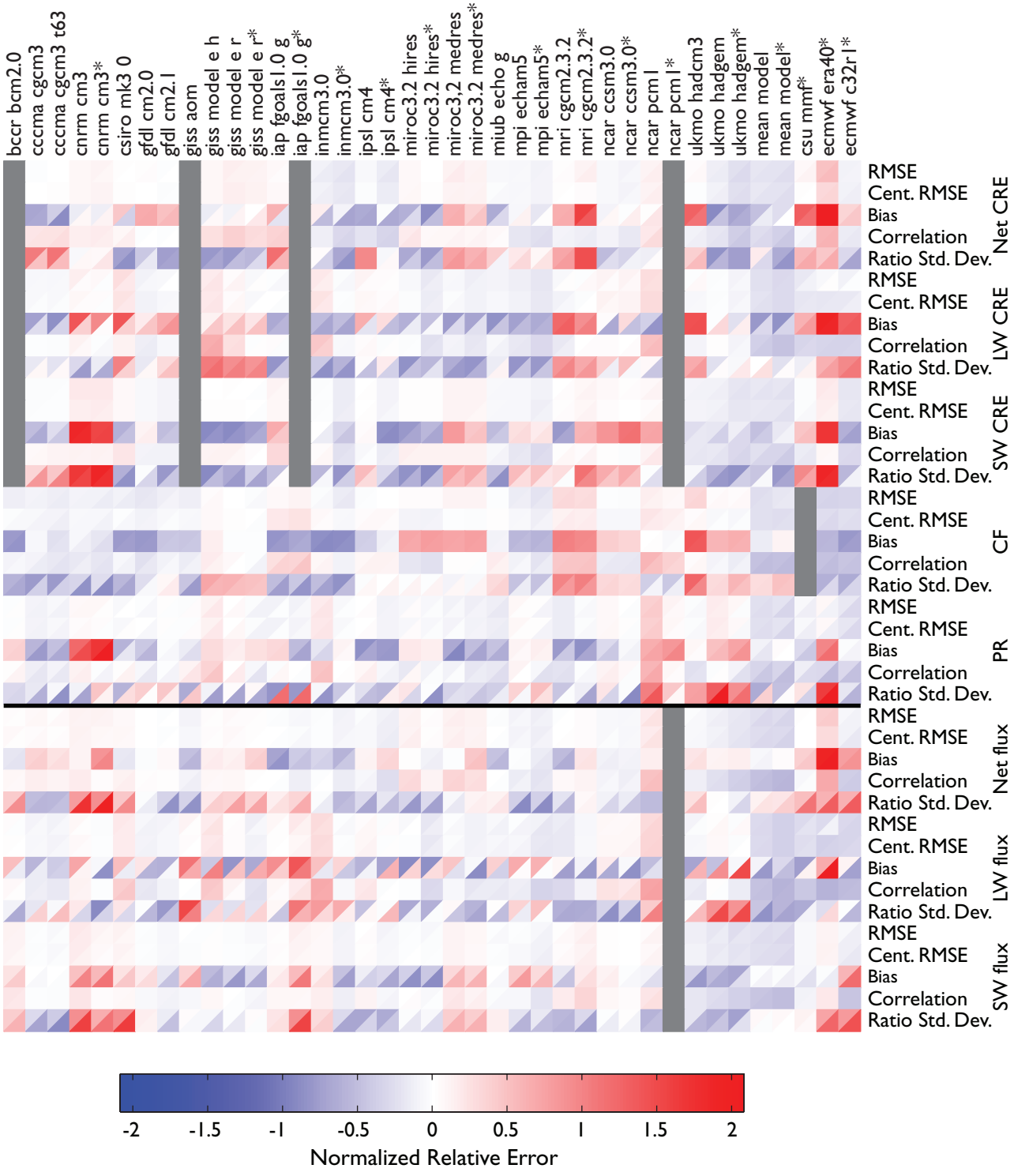
Institution	Country	Model name	Ensemble size - AMIP	Ensemble size - 20th C
		UKMO HadGem3	1	2
(None)		"IPCC mean model"	12	22
Colorado State University	USA	SuperCAM (1986-2000)	1	
European Centre for Medium-range Weather Forecasts	EU	ERA-40	1	
		ECMWF C32R1	1	

Figure 1. Portrait diagram showing the relative error in global model simulations of the annual cycle of top-of-atmosphere cloud effect, cloud fraction, surface precipitation rate, and (below the black line) top-of-atmosphere radiative fluxes. Each column corresponds to a model run (asterisks denote fixed-SST AMIP runs) and each column one of five statistical measures of agreement in one of eight physical parameters. White squares indicate that a model's score for a given metric is equal to the mean of that score across all the CMIP3 models (including the "mean model"); darker shades of blue indicate better-than-average scores and shades of red worse-than-average scores. Within each square the upper-left triangle denotes agreement with the primary observational data set and the lower-right triangle agreement with the alternate data set. Gray indicates that no data is available.

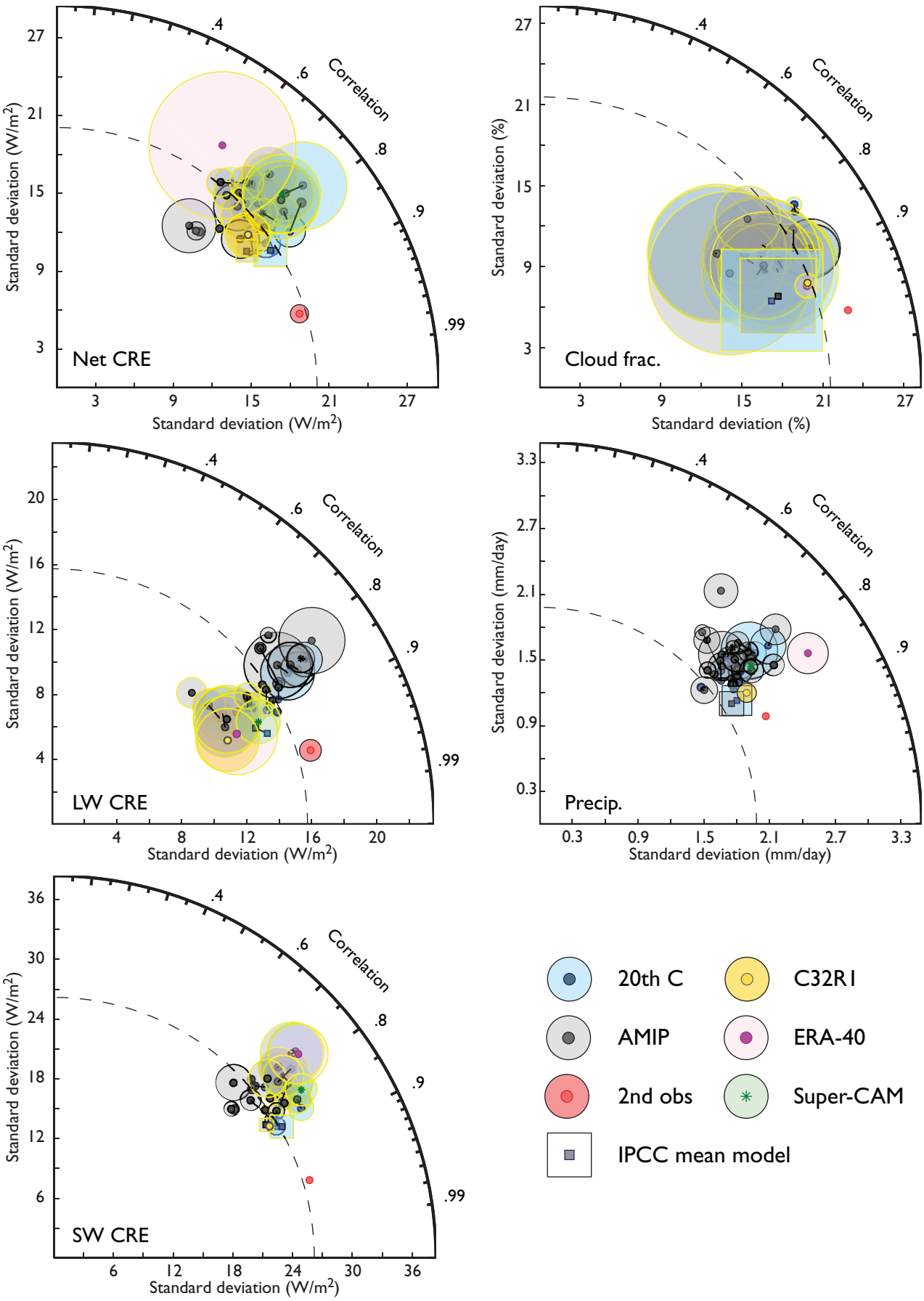
Figure 2. Summary Taylor diagrams showing errors in top-of-atmosphere radiative fluxes, top-of-atmosphere cloud forcing, cloud fraction, and surface precipitation rate computed over the mean annual cycle. Following Taylor (2001), the radial distance from the origin denotes the standard deviation of each data set (the primary observations are shown as a dashed radius) and the angular distance from the horizontal denotes the correlation coefficient R between each data set and the primary observations. The centered RMS error is indicated by the distance to the intersection of the dashed line and the x-axis with units and measures identical to the x-axis. Here the size of the symbol (diameter for circles, edge length for squares) indicates the mean bias and a yellow outline indicates a negative bias. Each diagram includes both the AMIP (blue) and 20th century (gray) runs; where a model has submitted both runs the points are joined by a line. The IPCC mean model, computed separately for the AMIP and 20th century ensembles, is shown as a square, while the ECMWF models are shown in pink (ERA40) and yellow (C32R1). The super-parameterized CAM is shown in green, and all three versions of the NCAR CAM are denoted with asterisks. Metrics for each quantity are computed with respect to the primary data set listed in Table 1; differences between the primary and secondary observational estimates are shown in red.

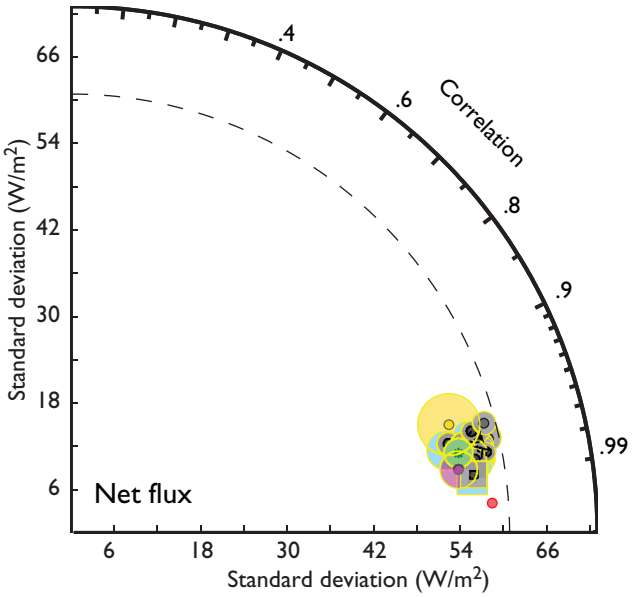
Figure 3. Taylor diagrams for net, longwave, and shortwave flux components at the top of the atmosphere evaluated against CERES. The color coding of the model points follows Figure 2. Agreement with the observations is much better than in Figure 2 because temporal and spatial variations in the TOA flux field are dominated by the well-understood pattern of insolation.

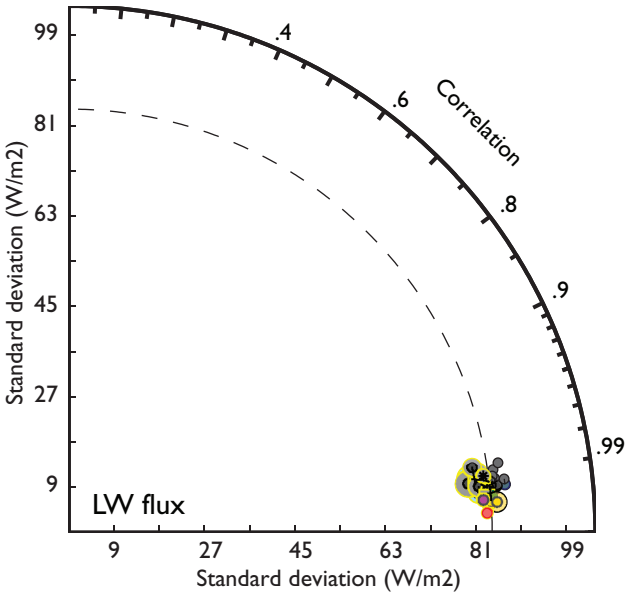
Figure 4. Distribution of errors across all metrics, relative to a single run of the MRI CGCM, for fields which have been averaged or spatially smoothed. The median of each distribution is shown as a horizontal bar and the inter-quartile range as a box. The whiskers extend to 1.5 times the inter-quartile range, with values beyond that marked with individual points. The sample size varies from column to column according to the number of model realizations used. All members of the MRI CGCM ensemble (the first column) show comparable errors, while the ensemble mean performs slightly better than the individual runs. Spatial smoothing improves many metric scores, but the most dramatic improvements come from averaging across multi-model ensembles, even when the ensemble size is no larger than the MRI CGCM ensemble.

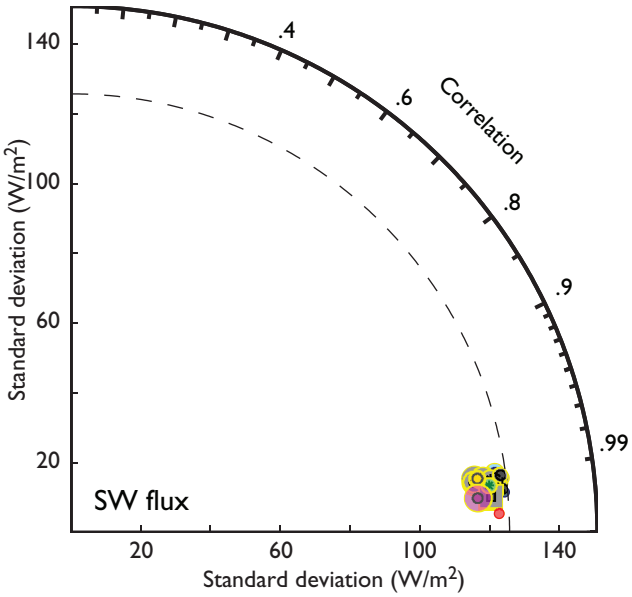


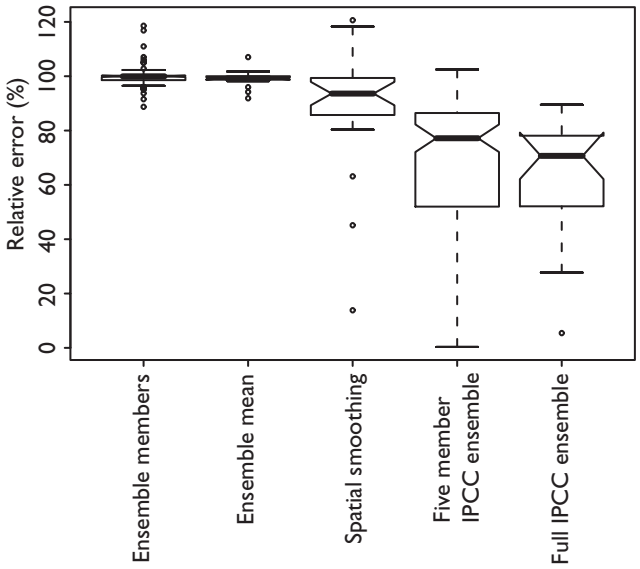
Pincus et al., Evaluating clouds and radiation in climate models: Figure 1


Net CRE
Cloud frac.
LW CRE
Precip.
SW CRE
Correlation
.4
.6
.8
.9
.99
Standard deviation (W/m2)
Standard deviation (%)
Standard deviation (mm/day)
20th C
AMIP
2nd obs
IPCC mean model
C32R1
ERA-40
Super-CAM


Net flux
Correlation
.4
.6
.8
.9
.99
Standard deviation (W/m2)
6
18
30
42
54
66


LW flux
Correlation
.4
.6
.8
.9
.99
Standard deviation (W/m2)
9
27
45
63
81
99


SW flux
Correlation
.4
.6
.8
.9
.99
Standard deviation (W/m2)
20
60
100
140



Pincus et al., Evaluating clouds and radiation in climate models: Figure 4